

Multiple Access for Small Packets Based on Precoding and Sparsity-Aware Detection

Ronggui Xie¹, Huarui Yin¹, Xiaohui Chen¹, and Zhengdao Wang²

¹Department of Electronic Engineering and Information Science, University of Science and Technology of China

²Department of Electrical and Computer Engineering, Iowa State University

Abstract

Modern mobile terminals often produce a large number of small data packets. For these packets, it is inefficient to follow the conventional medium access control protocols because of poor utilization of service resources. We propose a novel multiple access scheme that employs block-spreading based precoding at the transmitters and sparsity-aware detection schemes at the base station. The proposed scheme is well suited for the emerging massive multiple-input multiple-output (MIMO) systems, as well as conventional cellular systems with a small number of base-station antennas. The transmitters employ precoding in time domain to enable the simultaneous transmissions of many users, which could be even more than the number of receive antennas at the base station. The system is modeled as a linear system of equations with block-sparse unknowns. We first adopt the block orthogonal matching pursuit (BOMP) algorithm to recover the transmitted signals. We then develop an improved algorithm, named interference cancellation BOMP (ICBOMP), which takes advantage of error correction and detection coding to perform perfect interference cancellation during each iteration of BOMP algorithm. Conditions for guaranteed data recovery are identified. The simulation results demonstrate that the proposed scheme can accommodate more simultaneous transmissions than conventional schemes in typical small-packet transmission scenarios.

Index Terms

small packet, block-sparsity, compressive sensing, massive MIMO, BOMP, precoding, interference cancellation

Part of this work will be presented at the IEEE/CIC International Conference on Communications in China, 2014 [6].

I. INTRODUCTION

As intelligent terminals such as smart phones and tablets get more popular, they produce an increasing number of data packets of short lengths, to be delivered over a cellular network. Modern mobile applications that produce such small packets include instant messaging, social networking, and other services [1], [2]. Although the lengths of messages are relatively short, small packet services put great burden on the communication network. Two kinds of messages contribute to the traffic of small packets: one is the small packets of conversation produced by active users that occupy only a small percentage of the total online users [2]; the other is the signaling overheads needed to transmit these conversation packets [3].

In current wireless communication systems, a user follows the medium access control (MAC) protocols to obtain the service resources. Two flavors of MAC protocols are used in general: i) resource reservation based, and ii) collision resolution based. In the first kind, resources are preallocated to the users in a noncompetitive fashion. For small and random packets, the reservation-based approach is inefficient in resource utilization due to irregularity of the packets. In the collision-resolution based approaches, the terminals are allowed to access the resources in arbitrary order and when collision occurs, certain resolution mechanism is then employed. The collision-resolution based MAC can suffer from too many retransmissions due to frequent collisions.

In this paper, we propose a novel uplink small packet transmission scheme based on precoding at the transmitters and sparsity-aware detection at the receiver. The main motivation is to allow for a large number of users to transmit simultaneously, although each user may be transmitting only a small amount of data. Besides frame-level synchronization, no competition for resources or other coordinations are required. This saves the signaling overhead for collision resolution, and improves the resource utilization efficiency.

The contributions of our work are as follows:

- 1) *Block precoding and block-sparse system modeling*: We apply block precoding at each transmitter in time domain, and by considering the user activities, develop a block-sparse system model that takes full advantages of the structure of the signals to recover and is suitable for compressive-sensing based detection algorithms.
- 2) *Sparsity-aware detection algorithm*: We develop a interference cancellation (IC) based

block orthogonal matching pursuit (ICBOMP) algorithm. The algorithm improves upon the traditional BOMP algorithm by taking advantage of availability of error correction and detection, which is common in digital communications. By ICBOMP algorithm, not only do we achieve much better signal recovering accuracy but we also benefit in terms of less computational complexity. The price is slightly decreased rate due to coding.

- 3) *Signal recovery conditions:* We derive conditions for guaranteed signal recovery. The condition we require on the BOMP algorithm is milder than that in the related work in [20]. For ICBOMP algorithm, we give the conditions for perfect IC in each iteration. Furthermore, we characterize the data recovery condition from information theoretic point of view.

Thanks to the precoding operation and our sparsity-aware detection algorithms, our scheme enable the system to support more active users to be simultaneously served. The number of active users can be even more than the number of antennas at base station (BS). This is of great practical significance for networks offering small packet services to a large number of users.

Our proposed scheme is especially suitable for the so called massive multiple-input multiple-output (MIMO) systems [7]- [10]. In massive MIMO systems, the number of antennas at the BS can be more than the number of active single-antenna users that are simultaneously served. When the number of antennas at BS is large, the different propagation links from the users to the BS tend to be orthogonal, and the large amount of spatial degrees of freedom are useful for mitigating the effect of fast fading [8], [9]. Overall, massive MIMO technique provides higher data rate, better spectral and energy efficiencies [10].

Applications of compressive sensing (CS) to random MAC channels have been considered in [22]- [25]. In [22], CS based decoding scheme at the BS has been used for the multiuser detection task in asynchronous random access channels. A technique based on CS for meter reading in smart grid is proposed in [23], and its consideration is limited to single-antenna systems. Besides, a novel neighbor discovery method in wireless networks with Reed-Muller Codes has been proposed in [25], where CS technique is also adopted. All the referred works depend on the idea that the MAC channel is sparse, and all their works are classified to initial category of CS, where no structure property have been taken into account. This is one of the main distinctions that differentiate our work from the referred ones.

The rest of the paper is organized as follows. In Section II, the system model of block sparsity

are given. In Section III, we introduce the BOMP algorithm and its improved version ICBOMP algorithm to recover the transmitted signals. Guarantees for data recovery are presented in Section IV. Section V will present the numerical experiments that prove the effectiveness of our scheme. Afterwards, to invest the scheme with practical significance, some issues are discussed in Section VI. Finally, the conclusion will be presented in Section VII. To better organize the contents, we will relegate some of the proofs to the appendix.

Notation: Vectors and matrices are denoted by boldface lowercase and uppercase letters, respectively. The 2-norm of a vector \mathbf{v} is denoted as $\|\mathbf{v}\|_2$, and the 0-norm is given as $\|\mathbf{v}\|_0$. The inner product of two vectors \mathbf{v}_1 and \mathbf{v}_2 is denoted as $(\mathbf{v}_1, \mathbf{v}_2)$. The identity matrix of $d \times d$ dimension is denoted as \mathbf{I}_d . For a given matrix \mathbf{U} , its conjugate transpose, transpose, pseudo inverse, trace and rank are respectively denoted as \mathbf{U}^H , \mathbf{U}^T , \mathbf{U}^\dagger , $\text{Tr}\{\mathbf{U}\}$, $\text{rank}\{\mathbf{U}\}$, and the spectral norm of \mathbf{U} is given by $\|\mathbf{U}\|$. Operation $\text{vec}(\mathbf{U})$ denotes vectorizing \mathbf{U} by column stacking. For a subset $I \subset [N] := \{1, 2, \dots, N\}$ and matrix $\mathbf{U} := [\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_N]$ consisting of N sub-matrices (blocks), where each sub-matrix has equal dimensionality, \mathbf{U}_I denotes a sub-matrix of \mathbf{U} with block indices in I ; for a vector $\mathbf{v} := [\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_N^T]^T$, \mathbf{v}_I is similarly defined. For a set A , $|A|$ denotes its cardinality. For two sets A and B , $A \setminus B := A \cap B^c$ denotes the set difference. For a real number r , $|r|$ and $\text{Re}(r)$, and $\lfloor r \rfloor$ denote its absolute value, real part, and floor, respectively. Operation \otimes denotes the Kronecker product of two matrices.

II. SYSTEM MODEL

Consider an uplink system with N mobile users, each with a single antenna, and a BS with M antennas. When a terminal is admitted to the network, it becomes an online user. We assume that there are N_a active users, out of the total N online users, that have data to transmit. It is not required N_a be known a priori or $N_a < M$; actually, in practical systems N_a is usually unknown and it is possible that $N_a \gg M$.

We make the following further assumptions on the system considered.

- 1) The channels are block-fading: it remains constant for a certain duration and then changes independently.
- 2) The transmissions are in blocks and the users are synchronized at the block level. We assume that each frame of transmission consists of T symbols, which all fall within one channel coherent interval.

- 3) The users each have single antenna. There are multiple antennas at the BS.
- 4) The antennas at the BS, as well as the antennas among users, are uncorrelated and uncoupled.
- 5) The BS always has perfect channel state information (CSI) of online users.

Let $\mathbf{s}_n \in \mathbb{C}^{d \times 1}$ denotes the symbols to be transmitted by user n , with $d < T$. User n applies a precoding to \mathbf{s}_n to yield

$$\mathbf{x}_n = \mathbf{P}_n \mathbf{s}_n \quad (1)$$

where \mathbf{P}_n is a complex precoding matrix of size $T \times d$. The entries of \mathbf{x}_n are transmitted in T successive time slots. The received signals at all antennas within one frame can be written as

$$\mathbf{Y} = \sqrt{\rho_0} \sum_{n=1}^N \mathbf{h}_n \mathbf{x}_n^T + \mathbf{Z} = \sqrt{\rho_0} \sum_{n=1}^N \mathbf{h}_n \mathbf{s}_n^T \mathbf{P}_n^T + \mathbf{Z} \quad (2)$$

where ρ_0 is the signal to noise ratio (SNR) of the uplink, \mathbf{Y} is noisy measurement of size $M \times T$, $\mathbf{Z} \in \mathbb{C}^{M \times T}$ represents the additive noise, with i.i.d. circularly symmetric complex Gaussian distributed random entries of zero mean and unit variance, and $\mathbf{h}_n \in \mathbb{C}^{M \times 1}$ represents the channel coefficients from the user n to the base station, without loss of generality, let $h_{mn} \sim \mathcal{CN}(0, 1)$, $m = 1, 2, \dots, M$. Using the linear algebra identity $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B})$, we can rewrite the received signal as

$$\text{vec}(\mathbf{Y}) = \sqrt{\rho_0} \sum_{n=1}^N (\mathbf{P}_n \otimes \mathbf{h}_n) \mathbf{s}_n + \text{vec}(\mathbf{Z}) \quad (3)$$

Define $\mathbf{y} := \text{vec}(\mathbf{Y})$, $\mathbf{B}_n := (\mathbf{P}_n \otimes \mathbf{h}_n) / \sqrt{M}$ and $\mathbf{B} := [\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_N]$, $\mathbf{s} := [\mathbf{s}_1^T, \mathbf{s}_2^T, \dots, \mathbf{s}_N^T]^T$. Then we can write the model in (3) as

$$\mathbf{y} = \sqrt{\rho_0 M} \mathbf{B} \mathbf{s} + \mathbf{z} \quad (4)$$

In this formulation, we have assumed that all the users have messages of equal length d . This may not be the case in practice. We view d as the maximum length of the messages of all users within a frame. For the users whose message length is less than d , we assume their messages have been zero-padded to d before precoding. Also, for those users that are not active, we assume their transmitted symbols are all zeros.

Model (4) indicates that the signals to recover present the structure of block-sparsity where transmitted signals are only located in a small fraction of blocks and all other blocks are zeros.

The length of each block is d . We collect all the indices of blocks corresponding to active users to form a set I , with $|I| = N_a \leq K$, which means the unknown number of nonzero blocks (active users) is at most K . Our consideration is limited to case $MT < Nd$, where (4) represents an under-determined system. For case where $MT > Nd$, the receiver design is easier and the proposed method is also applicable. When precoding matrix \mathbf{P}_n is reasonably designed, matrix \mathbf{B} can meet the requirement for sensing matrix in CS, and this kind of \mathbf{P}_n is of wide range, for instance, Gaussian or Bernoulli matrix. Therefore, model (4) can be viewed as block sparsity model in CS [19], [20]. In CS, \mathbf{B} is referred as dictionary.

Remark 1: The following are several remarks on our precoding scheme:

- 1a) The precoding scheme is proposed because in reality, T is usually several times longer than the lengths of small packets. Also, the precoding scheme contributes to solving signal recovery problem in the situation where $N_a > M$.
- 1b) Each user knows its own precoding matrix and the BS knows all precoding matrices of all users.
- 1c) A basic requirement on the precoding matrix is that it should be full column rank, which is a requirement for data recovery. Additionally, in order to balance the power of every symbol of the messages before and after being precoded, each column of \mathbf{P}_n should be normalized to unit energy.
- 1d) Our precoding scheme is different from spreading schemes in [23], [24], where direct sequence spread spectrum (DSSS) is utilized for CS formulation.

III. ALGORITHMS FOR DATA RECOVERY

The past few years have also witnessed the research interest in CS [11]–[14]. Initial works in CS treat sparse weighting coefficients as just randomly located among possible positions in a vector. When structure of the sparse signal is exploited, for example block sparsity, it is possible to obtain better signal reconstruction performance and reduce the number of required measurements [15], [19]–[21].

We develop detection algorithms to be used at the BS in this section. We first apply known algorithm BOMP to our problem, and then further improve it by incorporating IC based on error correction and detection.

A. BOMP Algorithm

The main idea of BOMP algorithm is that, for each iteration, it chooses a block which has the maximum correlation with the residual signal, and after that, it will use the selected blocks to approximate the original signals by solving a least squares problem [20]. For later convenience, we present the details of BOMP algorithm as follows:

- 1) *Input*: Matrix \mathbf{B} , signal vector \mathbf{y} , ρ_0 , M , T and d .
- 2) *Parameter setting*: Maximum number of iterations K . Usually, $K \leq \lfloor \frac{MT}{d} \rfloor$. With K iterations, at most K active users can be identified.
- 3) *Initialization*: Index set $\Lambda_t = \emptyset$, basis function set $\Theta_0 = \emptyset$, residual signal $\mathbf{r}_0 = \mathbf{y}$, the number of iterations $t = 1$.
- 4) *Main iteration*: While $t \leq K$, do the following
 - 4a) Calculate the of correlation coefficients given by the residual signal with each column of \mathbf{B} , denoting as $\mathbf{B}^H \mathbf{r}_{t-1}$.
 - 4b) Find the index $\lambda_t \in \{1, 2, \dots, N\}$ of the block and the block unit \mathbf{B}_{λ_t} of \mathbf{B} , satisfying $\{\lambda_t, \mathbf{B}_{\lambda_t}\} = \arg \max_{j \in \{1, \dots, N\}} \|\mathbf{B}_j^H \mathbf{r}_{t-1}\|_2$.
 - 4c) Augment the index set and basis function set, and set the λ_t -th block of \mathbf{B} a zero sub-matrix

$$\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$$

$$\Theta_t = [\Theta_{t-1}, \mathbf{B}_{\lambda_t}]$$

$$\mathbf{B}_{\lambda_t} = \mathbf{0}$$

- 4d) Estimate the updated signals by least square (LS) algorithm

$$\bar{\mathbf{s}}_t = \arg \max_{\mathbf{s}_0} \|\mathbf{y} - \sqrt{\rho_0 M} \Theta_t \mathbf{s}_0\|_2$$

- 4e) Update the residual signals and the iteration number

$$\mathbf{r}_t = \mathbf{y} - \sqrt{\rho_0 M} \Theta_t \bar{\mathbf{s}}_t$$

$$t = t + 1$$

- 5) *Output*: $\bar{\mathbf{s}}_K$, which is the approximation most unlikely all-zero blocks corresponding to block index set Λ_K in the original block-sparse signal vector \mathbf{s} .

After we get the approximation of block-sparse coefficients, we can recover the original signals we want.

B. ICBOMP Algorithm

In addition to the block structure of the signals to recover, the amplitude of each modulated symbol has constant modulus, which could potentially be utilized for improved performance. More importantly, error correction coding is usually used for correcting demodulation errors due to noise and channel disturbances. Error detection codes such as cyclic redundancy check (CRC) codes can be utilized to indicate whether the decoded packets are indeed correct. Such detection cannot be perfect. However, for simplicity we will assume CRC detection is perfect, i.e., the probabilities of miss detection and false alarm are both zero.

Here we propose an improved algorithm ICBOMP which make use of channel coding and CRC to carry out perfect IC in each iteration of BOMP algorithm. The idea of IC can be found in [16]–[18].

Let \bar{s}_t^i denote each d -length block \bar{s}_t^i of \bar{s}_t , which is obtained by step 4d) in BOMP algorithm, $1 \leq i \leq t$. We assume that the transmitters have used certain error correction and detection scheme for each block signal, denoted by $[\tilde{\Lambda}_t, \tilde{s}_t] = \mathcal{D}(\bar{s}_t, d)$, in which \tilde{s}_t is the output of input \bar{s}_t , and each of its blocks is denoted by \tilde{s}_t^i ; $\tilde{\Lambda}_t$ denotes the index set for error-free blocks. For each block, it will go through one of the two distinct operations given by function \mathcal{D} :

- 1) If \bar{s}_t^i after operation of certain channel coding scheme is error-free, then output \tilde{s}_t^i is its corrected signal vector.
- 2) If \bar{s}_t^i after operation of certain channel coding scheme is not error-free, then $\tilde{s}_t^i = \bar{s}_t^i$.

In ICBOMP algorithm, most of the calculation processes remain the same as BOMP algorithm, and the difference occurs after signals have been updated by LS method in step 4d) of BOMP algorithm, turning residual signals updating steps to the following

$$[\tilde{\Lambda}_t, \tilde{s}_t] = \mathcal{D}(\bar{s}_t, d) \quad (5)$$

$$\mathbf{r}_t = \mathbf{y} - \sqrt{\rho_0 M} \mathbf{B}_{\Lambda_t} \tilde{s}_t \quad (6)$$

$$\Lambda_t = \Lambda_t \setminus \{\tilde{\Lambda}_t\} \quad (7)$$

$$t = t + 1 \quad (8)$$

From the above steps, we can see that, when some blocks of signals have been exactly recovered, ICBOMP algorithm regards them as interference signals to the following iterations and eliminates these signals, as well as their contributions to signal receiving model of (4). While

for the signal blocks with errors that cannot be corrected, ICBOMP algorithm leaves them as they were obtained through BOMP algorithm. In the case where no error-free block is available, ICBOMP behaves the same way as BOMP.

IV. DATA RECOVERY GUARANTEES

In this section, we will present conditions that guarantee data recovery. Before analyzing conditions for data recovery, some notation and definitions will be introduced first. From the definition of \mathbf{B} , we can see that each column of it is statistically normalized to one. Here we expand \mathbf{B} as

$$\mathbf{B} = [\underbrace{\mathbf{b}_1 \cdots \mathbf{b}_d}_{\mathbf{B}_1} \underbrace{\mathbf{b}_{d+1} \cdots \mathbf{b}_{2d}}_{\mathbf{B}_2} \cdots \underbrace{\mathbf{b}_{(N-1)d+1} \cdots \mathbf{b}_{Nd}}_{\mathbf{B}_N}] \quad (9)$$

As in [19], [20], we give the definitions of block-coherence as

$$\mu_{\mathbf{B}} := \frac{1}{d} \max_{i \neq j} \|\mathbf{B}_i^H \mathbf{B}_j\| \quad (10)$$

and sub-coherence as

$$\nu := \max_{1 \leq l \leq N} \max_{(l-1)d+1 \leq i \neq j \leq ld} |\mathbf{b}_i^H \mathbf{b}_j| \quad (11)$$

At the same time define

$$s_l := \min_{i \in I} \|\mathbf{s}_i\|_2, \quad s_u := \max_{i \in I} \|\mathbf{s}_i\|_2 \quad (12)$$

A. Data Recovery Conditions For BOMP Algorithm

The following theorem characterizes the block-sparse data recovery performance by BOMP algorithm.

Theorem 1. *Consider the block-sparse model in (4), suppose that condition*

$$\begin{aligned} \rho_0 M [1 - (d-1)\nu]^2 s_l^2 &> \tau^2 + \rho_0 M d \mu_{\mathbf{B}} \{2(N_a - 1)[1 + (d-1)\nu] + N_a^2 d \mu_{\mathbf{B}}\} s_l^2 \\ &+ 2\sqrt{\rho_0 M} \tau \{(2N_a - 1)d \mu_{\mathbf{B}} + [1 + (d-1)\nu]\} s_l \end{aligned} \quad (13)$$

is satisfied, then the BOMP algorithm identifies the correct support of signal vector \mathbf{s} and at the same time achieves a bounded error given by

$$\|\hat{\mathbf{s}} - \mathbf{s}\|_2^2 \leq \frac{K \tau^2}{[1 - (d-1)\nu - (K-1)d \mu_{\mathbf{B}}]^2 \rho_0 M} \quad (14)$$

where $\hat{\mathbf{s}}$ is the signal vector recovered by BOMP algorithm, $K \leq \lfloor \frac{MT}{d} \rfloor$ is the maximum number of iterations for BOMP algorithm, $1 - (d-1)\nu - (K-1)d\mu_{\mathbf{B}} > 0$ and $\tau = \max_{1 \leq j \leq N} \|\mathbf{B}_j^H \mathbf{z}\|_2$. For circularly symmetric complex Gaussian noise \mathbf{z} ,

$$P\{\tilde{\tau} \geq \|\mathbf{B}_j^H \mathbf{z}\|_2\} \geq 1 - e^{-\varsigma^2} \sum_{k=0}^{d-1} \frac{(\varsigma^2)^k}{k!} \quad (15)$$

where $\varsigma = \tilde{\tau} / \sqrt{1 + (d-1)\nu}$.

For a certain modulation constellation, suppose that each symbol's energy has been normalized, and the minimum distance between different symbols is l_{\min} , for example, $l_{\min} = \sqrt{2}$ for quadrature phase shift keying (QPSK) and $l_{\min} = 2$ for binary phase shift keying (BPSK), then by the bounded error in (14), we can conclude that the number of erroneously demodulated symbols is at most $N_e = \lfloor \|\hat{\mathbf{s}} - \mathbf{s}\|_2^2 / (l_{\min}/2)^2 \rfloor$. By now, we can present the expression of symbol error rate (SER) as

$$P_{\text{SER}} \leq \frac{N_e}{N_a d} \quad (16)$$

Remark 2: Since $T > d$ and $MT \gg d$, we can design orthogonal columns for precoding matrix \mathbf{P}_n of user n , $n = 1, 2, \dots, N$, then each block of dictionary \mathbf{B} is sub-matrix with orthogonal columns, meaning $\nu = 0$. On the other hand, we have $\tau \gg s_l$ when each nonzero element of \mathbf{s}_n satisfies a reasonable power constrain. Additionally, if $\mu_{\mathbf{B}} = 0$, then condition (13) can be simplified as $\rho_0 M s_l^2 > \tau^2 + 2\sqrt{\rho_0 M} \tau s_l \approx \tau^2$, which is milder when compared with [20, Theorem 5], which yields $\rho_0 M s_l^2 > 4\tau^2$ when applied to our scenario.

B. Conditions For Perfect IC In ICBOMP algorithm

Thanks to the error correction and detection, ICBOMP algorithm provides better performance than BOMP algorithm. In the case of perfect IC, the algorithm improves signal recovery quality and also reduces computational complexity. By eliminating the correctly decoded blocks and their contributions, the dimensionality of useful signals is reduced, which contributes to reducing the computations required in matrix inversion of LS algorithm. In the following, we present a theorem that characterizes the conditions for perfect IC in each iteration of ICBOMP algorithm.

When the first $i-1$ iterations and block selection in i -th iteration have been finished, suppose N_{ic}^{i-1} blocks of active users are already correctly recovered and cancelled by the previous $i-1$ iterations, and N_i blocks, whose indices are gathered to form a set I^i , will be substituted into

least square operation in i -th iteration, $N_{ic}^{i-1} + N_i = i$. Additionally, let set I_u^i contain the indices of unidentified active users and set I_{ic}^{i-1} contain the indices of active users that are already successfully recovered and cancelled. Then the following result holds

Theorem 2. *If conditions*

$$\begin{aligned} \rho_0 M[1 - (d-1)\nu_i]^2 s_{il}^2 &> \tau_i^2 + \rho_0 M d \mu_{iB} \{2(N_i - 1)[1 + (d-1)\nu_i] + N_i^2 d \mu_{iB}\} s_{il}^2 \\ &+ 2\sqrt{\rho_0 M} \tau_i \{(2N_i - 1)d \mu_{iB} + [1 + (d-1)\nu_i]\} s_{il} \end{aligned} \quad (17)$$

and

$$[1 - (d-1)\nu_i - (N_i - 1)d \mu_{iB}]^2 \rho_0 M t_c l_{\min}^2 \geq 4\tau_i^2 \quad (18)$$

are satisfied, then at least one block will be successfully recovered and cancelled in i -th iteration of ICBOMP algorithm. In (17) and (18), t_c is the number of bits that can be corrected by the channel coding scheme,

$$\tau_i = \max_{j \in \{[N] \setminus \{I_{ic}^{i-1} \cup I_u^i\}\}} \|\mathbf{B}_j^H (\sqrt{\rho_0 M} \mathbf{B}_{I_c^i} \mathbf{s}_{I_c^i} + \mathbf{z})\|_2 \quad (19)$$

and μ_{iB} , ν_i and s_{il} are respectively defined as μ_B , ν and s_l , and their definitions are only limited to users or active users in $\{[N] \setminus \{I_{ic}^{i-1} \cup I_u^i\}\}$.

In Theorem 2, we only considered the case where all blocks in I^i correspond to active users, based on the consideration that if signals of an active user cannot be successfully recovered when all the active users have been identified, they are less likely to be successfully recovered when non-active blocks begin to enter into the least square algorithm.

Proof of Theorem 2: Theorem 2 can be viewed as a corollary of Theorem 1.

When we consider each iteration in ICBOMP algorithm, by (17) which yields (35) in Appendix A, when applied to i -th iteration, the most likely active users will be identified. Treat the users in I_u^i and noise as perturbations for recovering the signal of users in I^i . With the same proof for Theorem 1, (17) can be verified. When (17) is satisfied, we achieve an error bounded by

$$\|\hat{\mathbf{s}}_{I^i} - \mathbf{s}_{I^i}\|_2^2 \leq \frac{N_i \tau_i^2}{[1 - (d-1)\nu_i - (N_i - 1)d \mu_{iB}]^2 \rho_0 M} \quad (20)$$

where we have utilized the knowledge $|I^i| = N_i$ which can be exactly obtained by ICBOMP algorithm. If condition

$$\|\hat{s}_{I^i} - s_{I^i}\|_2^2 \leq N_i t_c \left(\frac{l_{\min}}{2}\right)^2 \quad (21)$$

holds, then by correction of channel coding, at least one user's signals will surely be recovered without an error. By (21), (18) is obtained. ■

C. Condition From Information Theoretic Point Of View

From the BS's point of view, it is desirable to recover all the information conveyed by \mathbf{s} , including number of active users, exact indices of these active users, their transmitted information bits, etc.. When all the information are measured by bits, then The number of bits representing the indices of active users and signal bits of the transmitted messages are respectively $\log_2 \binom{N}{N_a}$ and $\sum_{i=1}^{N_a} b_i$. Assume all bits are generated with equal probability, and let S denote the set of bits needed to represent the total information, then

$$|S| \geq \log_2 \binom{N}{N_a} + \sum_{i=1}^{N_a} b_i \quad (22)$$

Remark 3: When the number of active users and lengths of the messages of active users are not prior known to BS, then the inequality in (22) is strictly established. Even when these two factors are prior known to BS, (22) still holds.

The following theorem roughly shows that the total bits of all information that can be recovered at the receiver can not exceed the capability of channel in a frame time.

Theorem 3. *Define p_e as the probability that some error has happened in the recovery of the information in set S , Then the following condition is necessary for the data recovery*

$$|S| \leq \frac{1}{1-p_e} [H(p_e) + \log_2 \det(\mathbf{I}_{MT} + \rho_0 \mathbf{B}_I \mathbf{B}_I^H)] \quad (23)$$

Proof of Theorem 3: Our proof of Theorem 3 mainly includes the properties of entropy, mutual information and Fano's Inequality [31].

We have

$$I(S; \mathbf{Y}, \mathbf{B}) = I(\mathbf{s}; \mathbf{Y}, \mathbf{B}) \quad (24)$$

$$= I(\mathbf{s}; \mathbf{B}) + I(\mathbf{s}; \mathbf{Y}|\mathbf{B}) \quad (25)$$

$$= I(\mathbf{s}; \mathbf{Y}|\mathbf{B}) \quad (26)$$

in which independence between \mathbf{s} and \mathbf{B} are utilized, which means $I(\mathbf{s}; \mathbf{B}) = 0$. By property

$$H(S|\mathbf{Y}, \mathbf{B}) = H(S) - I(S; \mathbf{Y}, \mathbf{B}) \quad (27)$$

$$= H(S) - I(\mathbf{s}; \mathbf{Y}|\mathbf{B}) \quad (28)$$

$$\geq |S| - C \quad (29)$$

where $C = \max_{p(\mathbf{x})} I(\mathbf{s}; \mathbf{Y}|\mathbf{B})$ is the maximum mutual information (channel capacity). On the other hand, by Fano's Inequality

$$H(S|\mathbf{Y}, \mathbf{B}) \leq H(p_e) + p_e|S| \quad (30)$$

combining (27) and (30), and using the well-known result $C = \log_2 \det(\mathbf{I}_{MT} + \rho_0 \mathbf{B}_I \mathbf{B}_I^H)$ [26], the desired inequality follows. \blacksquare

Remark 4: It can be seen that when $p_e \rightarrow 0$, the right hand side of the inequality converges to C . It means that we cannot hope to decode correctly information (including all information useful to the BS) at a rate higher than the capacity of the channel, assuming the availability of the information of the set of active users and their channels.

V. NUMERICAL RESULTS

The experimental studies for verifying the proposed scheme are presented in this section. In all simulations, the channel response matrix is i.i.d. Gaussian matrix of complex values and the N_a active users are chosen uniformly at random among all N online users. As for the block-sparse data vectors to be transmitted, we assume QPSK for data modulation. All results are presented with symbol error rate (SER) and frame error rate (FER) versus E_s/N_0 , where E_s is the symbol energy, N_0 is the noise spectral density. In the simulations with BOMP algorithm, we do not set the number of antennas to a large value, say one hundred or more, for the sake of simplicity. Besides, we will choose the frame length to be a multiple of the maximum length of short messages.

A. Influences Of Different Parameters

In the first experiments, we have the simulations of BOMP algorithm to check the influences of different parameters. We assume that all the messages have the same length d , and we

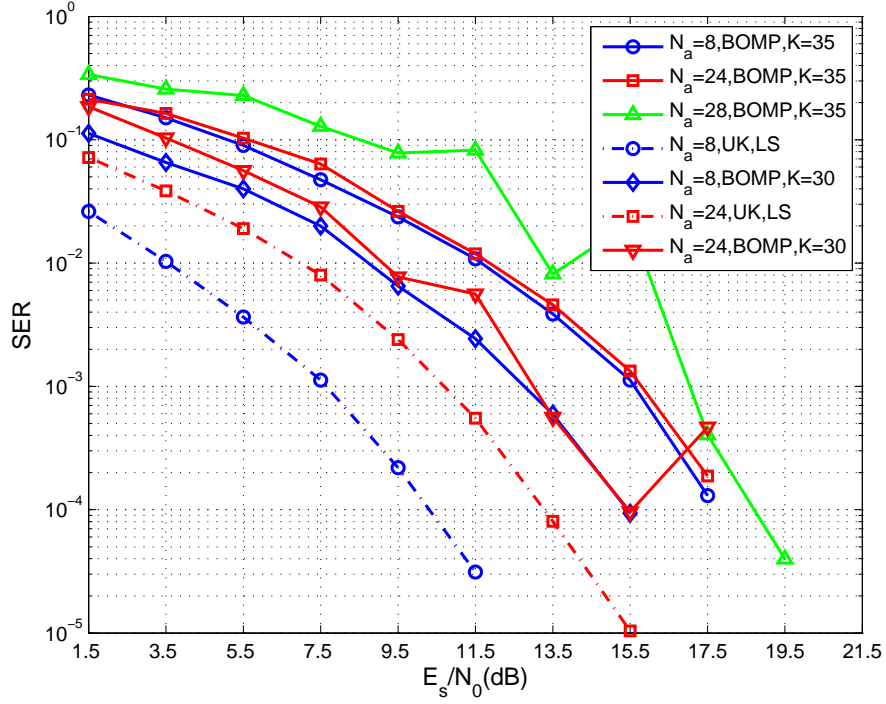


Fig. 1. symbol recovery with 8 antennas at BS

simply design \mathbf{P}_n a random matrix with ($v = 0$) or without ($v \neq 0$) orthogonal columns, $n = 1, 2, \dots, N$.

In our simulation, the SER is computed as follows: when a demodulated signal of an active user is different from its original signal, we claim a symbol error; if an active user is not identified, then all d symbol of that user are treated as erroneous.

Test Case 1: Figure 1 shows the performance of the proposed scheme with 8 antennas at BS, where K is the number of iterations for BOMP algorithm. Other parameters are given as $(N, d, T, v) = (80, 200, 1000, 0)$. The results indicate that, the SER increases when the number of active users becomes larger. For case where number of iterations is 35, when the number of active users is lower than a certain number, say 24 in our results, the SER is basically independent of the number of active users. Besides, we have observed that out of 35 iterations, in most cases the N_a ($N_a < 24$) active users can be successfully identified.

Also in Figure 1, we give results when less number of iterations is set for BOMP algorithm. When there are not too many active users, such as 8 users, fewer iterations are needed. But

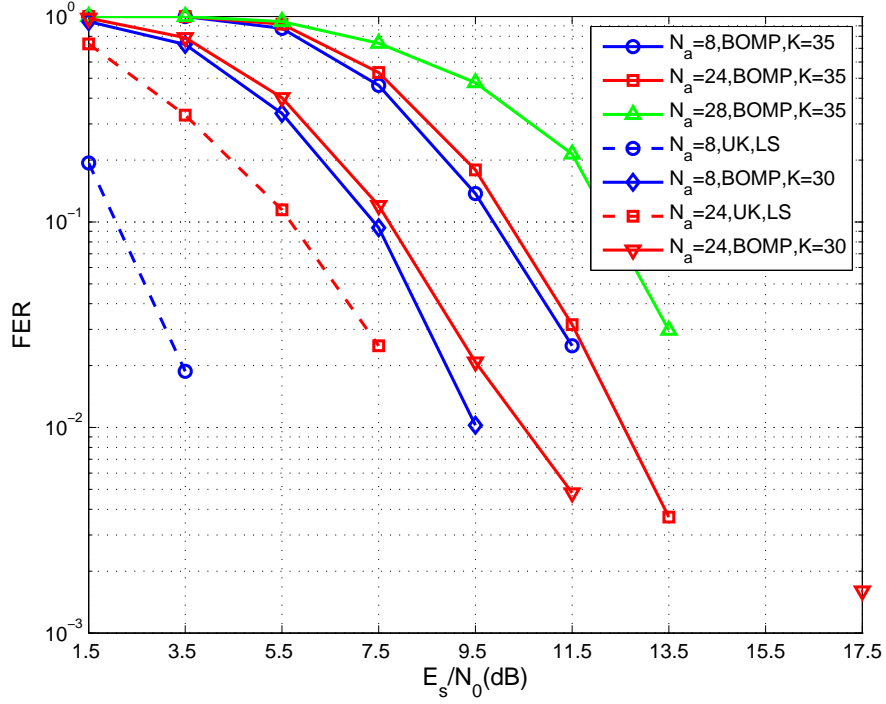


Fig. 2. frame recovery with 8 antennas at BS

for 24 active users, 30 iterations are not enough to include all the active users. We also plotted two curves of reference in dotted lines. Both curves were obtained under the conditions that the actual active users are already known (users known, UK) and picked out. From the results, we can conclude that the number of non-active users has a great influence on the performance, even greater than that of active users.

Test Case 2: Corresponding to Figure 1 in SER, the FER is depicted Figure 2 with the same settings. In our simulation, the FER is computed as follows: when more than 8 bits in a message are demodulated in error, we claim a frame error. If the bit errors are equal to or less than 8, we hypothesize that they can be detected and corrected by the channel coding schemes. The same trend in FER performance can be observed as SER. When the E_s/N_0 exceeds a certain threshold, the FER will be negligible.

The normalized throughput is defined as $(1 - P_{\text{FER}})N_a d / (MT)$, where $(1 - P_{\text{FER}})N_a$ is the maximum number of allowed active users in our scheme, P_{FER} is the value of FER; and MT/d is the maximum number of users that can be served when all time slots of a frame are effectively

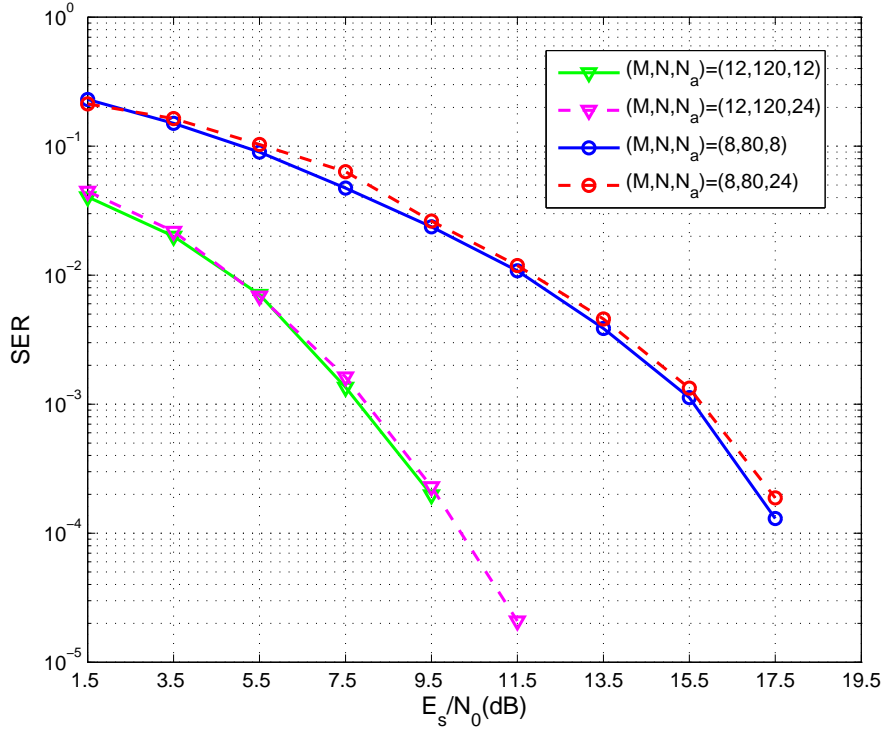


Fig. 3. symbol recovery with different numbers of antennas at BS

used for data transmission, which is 40 under the given parameters. If 24 active users are allowed to be simultaneously served, the throughput will reach 60% of maximum possible. In contrast, in conventional random access protocols, if we treat the signaling messages (such as request-to-send (RTS) signaling [4], [5]) also as small packets, the system throughput will be no more than 60%. Furthermore, if collision happens, which is often the case, the throughput will decrease a step further. Therefore, our scheme will greatly improve the system throughput compared to conventional schemes.

Test Case 3: In Figure 3, we compare the performance when BS are equipped with different numbers of antennas. Other parameters are given as $(d, T, K, v) = (200, 1000, 35, 0)$. The results show that, when the number of antennas M increases, the SER performance becomes remarkably better and a higher ratio N_a/M can be accepted. On the other hand, a big performance gap between 8 antennas and 12 antennas at BS is observed. More antennas at BS allows a larger number of iterations for BOMP algorithm to accommodate more active users, and the big

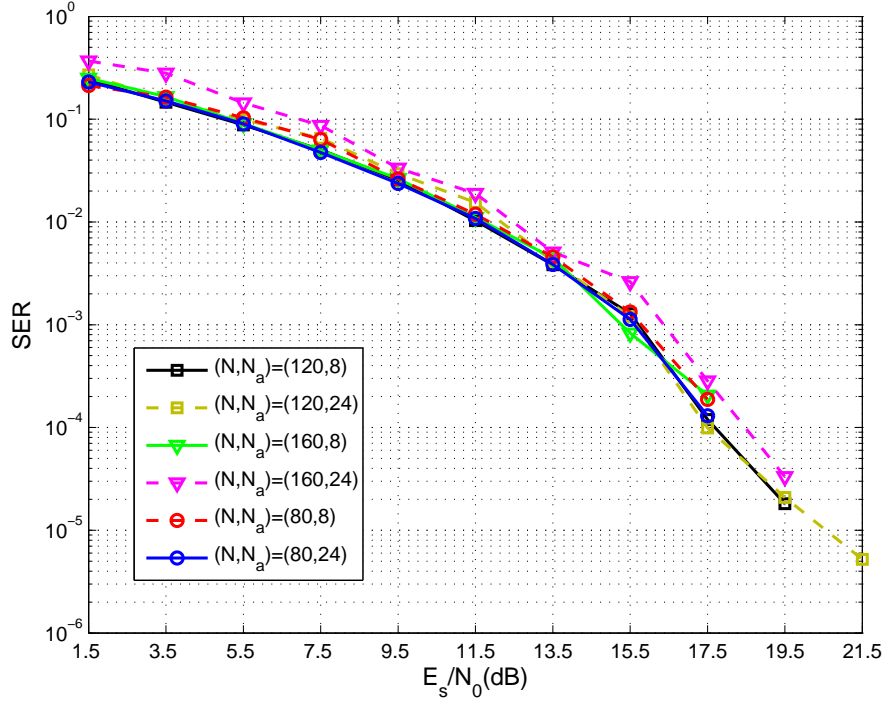


Fig. 4. symbol recovery with different numbers of online users

performance gap appears when we set both cases the same number 35 of iterations.

Test Case 4: From Figure 4 with parameters $(M, d, T, K, v) = (8, 200, 1000, 35, 0)$, we can see that when the number of active users is fixed, the SER increases as the number of online users increases, but the performance degradation is rather small, even when the number of online users has been doubled, nearly no more than 1dB degradation can be observed for 24 active users. By Theorem 3, the number of online users is not the dominant factor to affect the performance under the given parameters.

Test Case 5: Figure 5 depicts the performance when frame lengths are different, respectively for $T = 4d$, $T = 5d$ and $T = 6d$. Other parameters are given as $(M, N, d, v) = (8, 80, 200, 0)$. The number of iterations K is set to 28, 35 and 42, respectively. The results show that the longer the frame length is, the better performance, and hence the more users that can be simultaneously served. However, affected by the normalization of columns in precoding matrix, even when the length of frame grows, the benefits diminish. This phenomenon will be observed when parameters

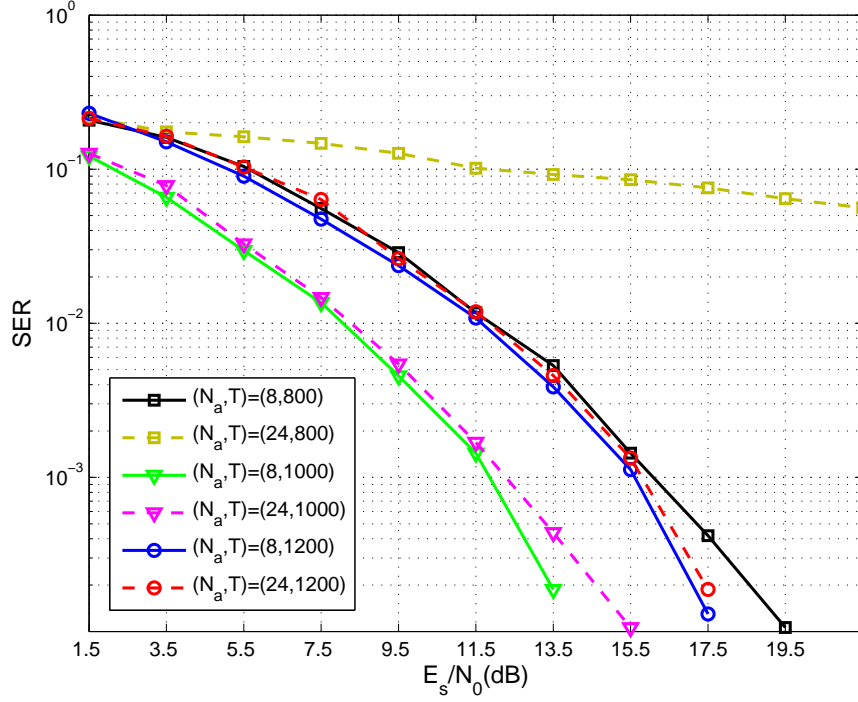


Fig. 5. symbol recovery with different lengths of frames

are chosen to ensure that $MT/(dK)$ is a constant.

Test Case 6: In Figure 6, we investigate the effect of orthogonality of the block of the dictionary on the performance. Other parameters are given as $(M, N, d, T, K) = (8, 80, 200, 1000, 35)$. As we can see, for column non-orthogonal case, nearly 1dB performance degradation is observed when compared with that of column orthogonal condition.

The results of our theoretical analysis of Theorem 1 for the case where $v = 0$ and all messages have same length are shown in Table I, giving minimum number of active users that can be simultaneously served. Obviously, the bounded minimum number of active users is pessimistic when compared with our simulation results. At the same time, such pessimistic result can also be seen in the theoretical analysis for SER, for condition $1 - (d - 1)\nu - (K - 1)d\mu_B > 0$ can not always be satisfied.

Remark 5: In all above simulations, we have set d as the length of all messages. In practice, this may not be the case. In fact, when different lengths for messages exist and the number of active users is large, it has some slight performance degradation. For example, similar result can be

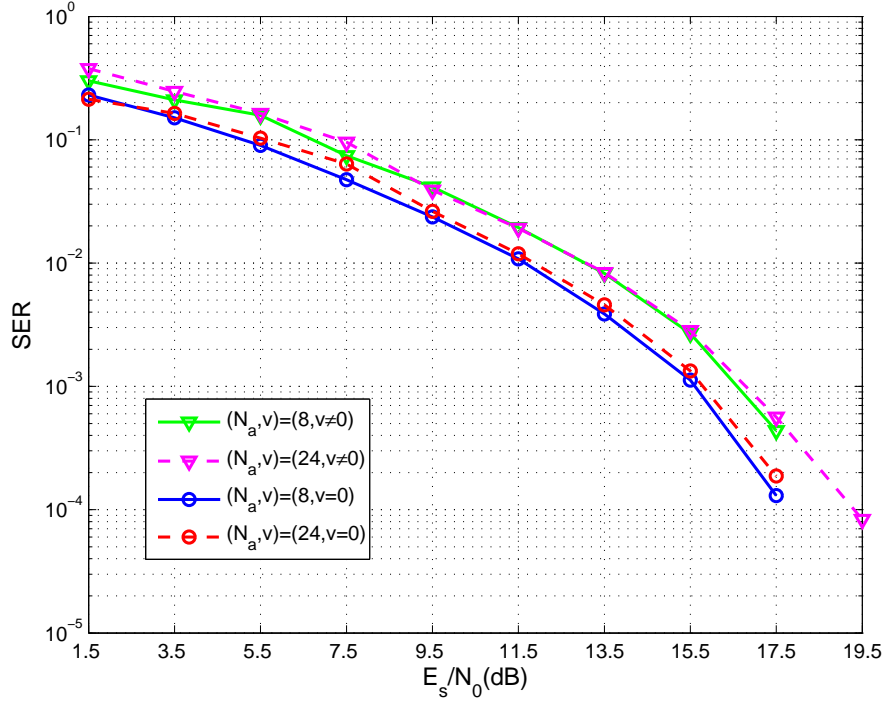


Fig. 6. symbol recovery with or without column orthogonal blocks

observed when there are 24 active users and length of each message is uniformly distributed in the interval from 50 to 200. On the other hand, performance gap between $d = 100$ and $d = 200$ is barely visible.

B. Performance Of ICBOMP Algorithm

In this part, we will apply ICBOMP algorithm to recover the transmitted signals.

Test Case 7: The results of FER for 8 antennas at BS are depicted in Figure 7, and other parameters are given as $(N, d, T, v) = (80, 200, 1000, 0)$. For ICBOMP algorithm, we choose the channel coding scheme that is capable of correcting at most 8 bits for message of 200 symbols, e.g., a shortened BCH(472,400), which enjoys a relatively high code rate. When the error of a message is beyond correction, a frame error is declared.

Compared with Figure 1 achieved by BOMP algorithm, Figure 7 shows that ICBOMP algorithm can greatly improve the recovering performance, and more active users can be served simultaneously, say 30 or 32. Thus the throughput will be increased a step further, reaching 75%

TABLE I
THEORETICAL ANALYSIS ($v = 0$)

M	N	d	T	$\langle s_l \rangle \langle s_u \rangle$	β	$\mu_{\mathbf{B}}$	τ	$N_{a.lower}(E_s/N_0)$		
								0dB	10dB	15dB
8	80	200	1000	14.14	2	0.0035	15.00	0	1	1
50	500	200	1000	14.14	2	0.0019	15.00	1	1	1
100	1000	200	1000	14.14	2	0.0014	15.00	1	2	2
8	80	200	1000	14.14	2	0.0035	14.20	0	1	1
50	500	200	1000	14.14	2	0.0019	14.20	1	1	1
100	1000	200	1000	14.14	2	0.0014	14.20	1	2	2
8	80	100	500	10	2	0.0066	15.00	0	1	1
50	500	100	500	10	2	0.0037	15.00	1	1	1
100	1000	100	500	10	2	0.0030	15.00	1	1	1

or 80% of the maximum possible. When ICBOMP algorithm is applied, the result we obtain for 24 active users with 35 iterations is almost identical to that achieved by 30 iterations, and we just depict one of them for clarity. Therefore, ICBOMP algorithm can also narrow the performance gap between different iteration numbers. It is also noticed that the performance gap between different numbers of active users have been widened.

Also, we included several curves in dotted lines to show the performance when actual active users are already picked out and no non-active users are chosen and interference-cancellation minimum mean-squared error (IC-MMSE) receiver is used. The IC-MMSE receiver performs iteratively MMSE decoding and perfect IC until more iterations no longer benefit. Perfect IC in IC-MMSE receiver is operated as in the ICBOMP algorithm. It shows that our ICBOMP receiver achieves a little worse performance than the IC-MMSE receiver, about 1dB performance degradation for 24 active users and 2dB performance degradation for 28 active users. Although for ICBOMP receiver, performance degradation exists when compared with IC-MMSE receiver, it is still highly competitive, since it requires no knowledge about active users. In fact, the ICBOMP receiver is able to achieve better performance than MMSE receiver when the number of active users is no more than 30.

Test Case 8: For massive MIMO, 8 or 12 antennas at BS are not enough. In the following, we present the FER performance with 32 and 64 antennas at BS in Figure 8, and other parameters

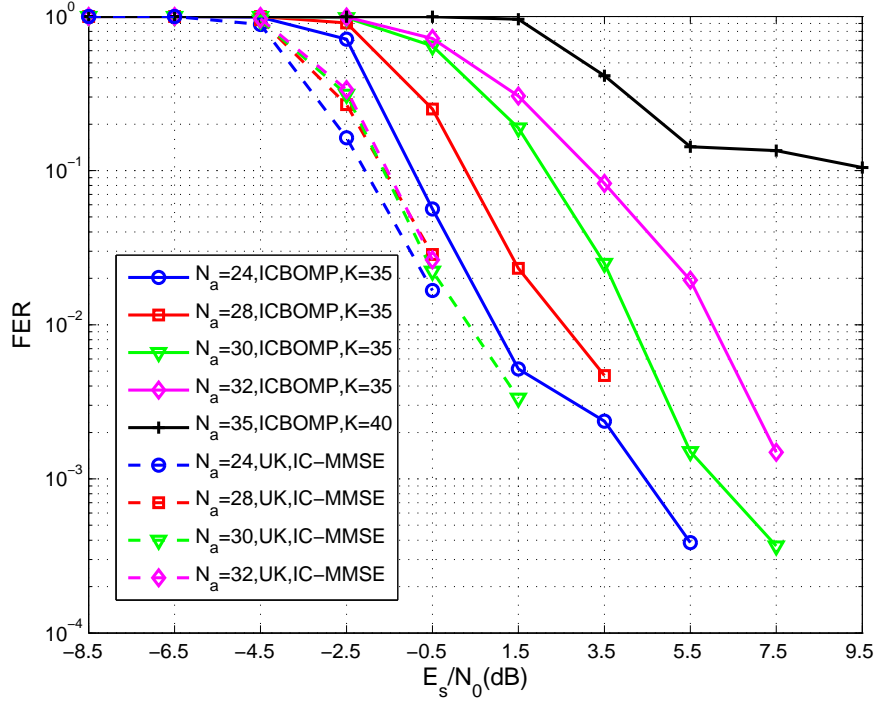


Fig. 7. Frame recovery with 8 antennas at BS by ICBOMP algorithm

are given as $(d, T, v) = (31, 155, 0)$. To ease the computational burden, we set the length for each message to 31. Channel coding scheme is assumed such that up to 1 bit of error can be corrected; e.g., a shortened BCH(69,62) could be used. When the number of errors in a message is larger than the designed error correction capability, a frame error is declared. The results show that, with more antennas, great improvements in performance are observed, and many more active users and online users can be accommodated.

VI. DISCUSSIONS

In this section, we discuss a few issues related to the design of our proposed scheme and its practical significance.

A. Dictionary Design

By Theorem 1, the smaller the block-coherence μ_B and ν are, the better performance we can achieve, and thus the more active users the model can simultaneously accommodate. The

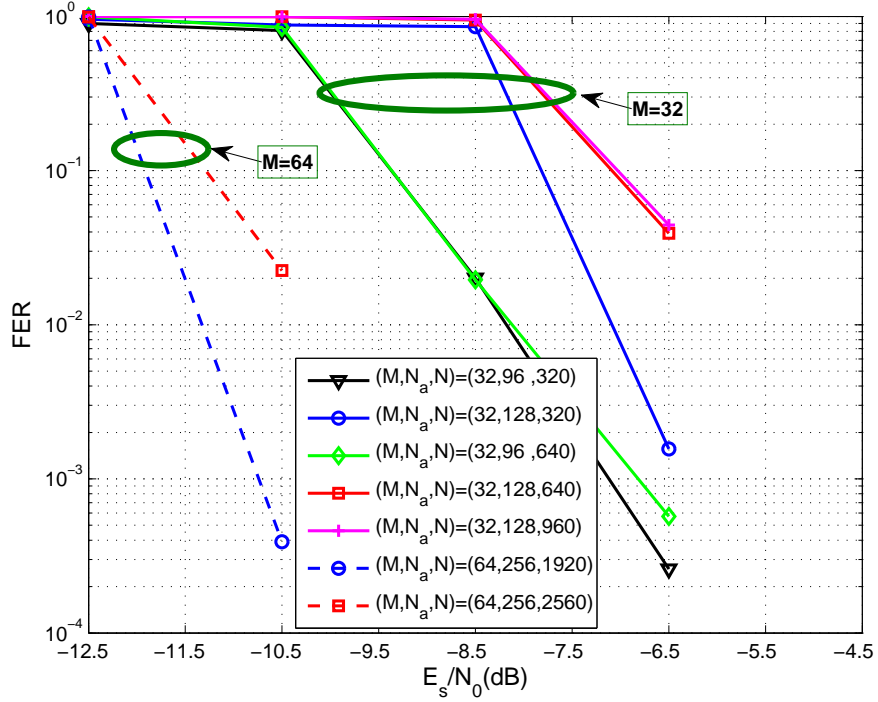


Fig. 8. frame recovery with 32 and 64 antennas at BS by ICBOMP algorithm

authors in [19] had a discussion about the design of the dictionary that can lead to significant improvement in data recovery in block-sparse model. In our model, only the precoding matrix is up to our design.

In our scheme, when the number of active users is more than the number of antennas at BS, the channel vectors among users are correlated, even with massive MIMO technique. However, by our precoding scheme, correlations among columns in \mathbf{B} can be smaller than correlations among channel vectors of different users, which means that block-coherence $\mu_{\mathbf{B}}$ can still be rather small, as long as precoding is well designed.

B. Application To Asynchronous Setting

Random MAC channel is usually asynchronous [22], while in our model, frame synchronization is assumed. In fact, our model can also be adjusted to quasi-synchronous setting, as long as the maximum asynchronism is known to the receiver. In such scenario, we can handle the problem by lengthening the maximum length d with addition of the maximum asynchronous

level, leading to a slight adjustment in the length of each precoding sequence. Readers can refer to [22] for more details.

C. Sparsity Level Selection

In our model, since we do not know exactly the number of active users a priori, a large number of iterations are usually needed for BOMP or ICBOMP algorithm. When there are not too many active users, however, unnecessary iterations increases computational cost. For example, our statistical results indicate that, when $E_s/N_0 \geq 10\text{dB}$ and with the same conditions for simulation of 8 active users in Figure 1, 12 iterations for BOMP algorithm will correctly identify all the active users with a probability exceeding 99.9%. And for ICBOMP algorithm, even less iterations and higher probability can be anticipated. To address this problem, some methods for sparsity level selection will work, such as sparsity adaptive matching pursuit in [30]. When the number of active users is large, sparsity level identification may not bring too much benefit.

D. Message Segmentation

Messages of small packet can be segmented into shorter parts further, and each part can be transmitted with our scheme. When this method is adopted, we not only alleviate the computational complexity, but also ease the requirement for coherent time of channel. Besides, we can decrease the length uncertainty for each segmented part, and shorten the transmission duration for small packets whose lengths are short. More importantly, such approach may even be adopted to data transmission for messages that are not belonging to small packet.

VII. CONCLUSION

In this paper, we proposed an uplink data transmission scheme for small packets. The proposed scheme combines the techniques of block precoding and sparsity-aware detection. It is especially suitable for system with a large number of antennas at the base station. Under the assumption that the BS has perfect CSI of every online user, we developed a block-sparse system model and adopted BOMP algorithm and its improved version ICBOMP algorithm to recover the transmitted data. With the ICBOMP algorithm which utilizes the function of error correction and detection

coding to perform perfect IC in each iteration, an significant performance improvement was observed.

The transmission scheme considered in this paper is applicable to future wireless communication system. The reason is that small packets play a more and more important role in the data traffic due to the wide usage of intelligent terminals. The overall throughput of such a system is currently hampered by small packets because of the heavy signaling overhead. Our scheme will greatly reduce the signaling overhead and improve the throughput of such systems.

APPENDIX A

PROOF OF THEOREM 1

We first present a few results and lemmas that are useful for the proof of Theorem 1. Our proof of Theorem 1 follows along the lines of [20].

First of all, we present two useful results that have been obtained in some literature cites.

Result 1. [20, Lemma 1] *Given the dictionary \mathbf{B} of normalized columns with block-coherence $\mu_{\mathbf{B}}$ and sub-coherence ν , it holds that*

$$\max_{i \neq j} \|\mathbf{B}_i^H \mathbf{B}_j\| \leq d\mu_{\mathbf{B}} \quad (31)$$

and

$$1 - (d - 1)\nu \leq \|\mathbf{B}_i^H \mathbf{B}_i\| \leq 1 + (d - 1)\nu \quad (32)$$

Provided that $1 - (d - 1)\nu - (K - 1)d\mu_{\mathbf{B}} > 0$ and $|I| \leq K$, it holds that

$$\|(\mathbf{B}_I^H \mathbf{B}_I)^{-1}\| \leq [1 - (d - 1)\nu - (K - 1)d\mu_{\mathbf{B}}]^{-1} \quad (33)$$

Result 2. [29, §10.2] *Denote $m \times n$ matrices \mathbf{A} and \mathbf{C} , whose singular values are $\sigma_1 \geq \dots \geq \sigma_n$, $\gamma_1 \geq \dots \geq \gamma_n$, respectively, then*

$$-\sum_{i=1}^n \sigma_i \gamma_i \leq \text{Re}[\text{Tr}(\mathbf{A}\mathbf{C}^H)] \leq \sum_{i=1}^n \sigma_i \gamma_i \quad (34)$$

In the following, two lemmas will be given.

Lemma 1. Consider the block-sparse model in model (4) and the condition $\tau = \max_{1 \leq j \leq N} \|\mathbf{B}_j^H \mathbf{z}\|_2$. Provided that

$$\begin{aligned} \rho_0 M [1 - (d-1)\nu]^2 s_u^2 &> \tau^2 + \rho_0 M (d\mu_{\mathbf{B}})^2 N_a^2 s_u^2 + 2\rho_0 M d\mu_{\mathbf{B}} \{(N_a - 1)[1 + (d-1)\nu]\} s_l^2 \\ &\quad + 2\sqrt{\rho_0 M} \tau \{(N_a - 1)d\mu_{\mathbf{B}} + [1 + (d-1)\nu]\} s_l + 2\sqrt{\rho_0 M} N_a d\mu_{\mathbf{B}} \tau s_u \end{aligned} \quad (35)$$

it holds that

$$\max_{j \in I} \|\mathbf{B}_j^H \mathbf{y}\|_2 > \max_{j \notin I} \|\mathbf{B}_j^H \mathbf{y}\|_2 \quad (36)$$

Proof of Lemma 1: Note that

$$\begin{aligned} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 &= \mathbf{y}^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{y} \\ &= \text{Tr}\{\mathbf{B}_j^H \mathbf{y} \mathbf{y}^H \mathbf{B}_j\} \\ &= \text{Tr} \left\{ \mathbf{B}_j^H \left[\sqrt{\rho_0 M} \left(\sum_{i \in I} \mathbf{B}_i \mathbf{s}_i \right) + \mathbf{z} \right] \left[\sqrt{\rho_0 M} \left(\sum_{i \in I} \mathbf{B}_i \mathbf{s}_i \right) + \mathbf{z} \right]^H \mathbf{B}_j \right\} \end{aligned} \quad (37)$$

Then we have

$$\begin{aligned} \max_{j \notin I} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 &= \max_{j \notin I} \text{Tr}(\mathbf{B}_j^H \mathbf{y} \mathbf{y}^H \mathbf{B}_j) \\ &= \max_{j \notin I} \rho_0 M \cdot \text{Tr} \left[\left(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \right) \left(\sum_{i \in I} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j \right) \right] \\ &\quad + \max_{j \notin I} 2\sqrt{\rho_0 M} \cdot \text{Re} \left\{ \text{Tr} \left[\left(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \right) \mathbf{z}^H \mathbf{B}_j \right] \right\} + \max_{j \notin I} \text{Tr}(\mathbf{B}_j^H \mathbf{z} \mathbf{z}^H \mathbf{B}_j) \end{aligned} \quad (38)$$

Note that for vectors $\tilde{\mathbf{x}} \in \mathbb{C}^{n \times 1}$ and $\tilde{\mathbf{y}} \in \mathbb{C}^{n \times 1}$, the matrix $\tilde{\mathbf{x}} \tilde{\mathbf{y}}^H$ at most has one nonzero singular value which equals to the absolute value of $\tilde{\mathbf{x}}^H \tilde{\mathbf{y}}$. In addition, for any matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, matrix $\tilde{\mathbf{x}} \tilde{\mathbf{y}}^H \mathbf{A}$ also has at most one nonzero singular value which equals to the absolute value of $\tilde{\mathbf{x}}^H \mathbf{A}^H \tilde{\mathbf{y}}$, and it holds that $\|\tilde{\mathbf{x}} \tilde{\mathbf{y}}^H \mathbf{A}\| = |\tilde{\mathbf{x}}^H \mathbf{A}^H \tilde{\mathbf{y}}| = |(\tilde{\mathbf{x}}, \mathbf{A}^H \tilde{\mathbf{y}})| \leq \|\tilde{\mathbf{x}}\|_2 \|\mathbf{A}^H \tilde{\mathbf{y}}\|_2 \leq \|\mathbf{A}\| \|\tilde{\mathbf{x}}\|_2 \|\tilde{\mathbf{y}}\|_2$. Together with Result 1 and Result 2, we have

$$\max_{j \notin I} \text{Tr} \left\{ \left(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \right) \left(\sum_{i \in I} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j \right) \right\} \quad (39)$$

$$\begin{aligned} &\leq N_a \{ \max_{i \neq j} \text{Tr} \{ (\mathbf{B}_j^H \mathbf{B}_i) (\mathbf{s}_i \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j) \} \} + (N_a^2 - N_a) \{ \max_{i \neq j \neq r} \text{Tr} \{ (\mathbf{B}_j^H \mathbf{B}_i) (\mathbf{s}_i \mathbf{s}_r^H \mathbf{B}_r^H \mathbf{B}_j) \} \} \\ &\leq N_a \{ \max_{i \neq j} \|\mathbf{B}_j^H \mathbf{B}_i\| \|\mathbf{s}_i \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j\| \} + (N_a^2 - N_a) \{ \max_{i \neq j \neq r} \|\mathbf{B}_j^H \mathbf{B}_i\| \|\mathbf{s}_i \mathbf{s}_r^H \mathbf{B}_r^H \mathbf{B}_j\| \} \end{aligned} \quad (40)$$

$$\leq N_a \{d\mu_{\mathbf{B}} d\mu_{\mathbf{B}} s_u^2\} + (N_a^2 - N_a) \{d\mu_{\mathbf{B}} d\mu_{\mathbf{B}} s_u^2\} \quad (41)$$

$$= N_a^2 (d\mu_{\mathbf{B}})^2 s_u^2 \quad (42)$$

where we have used the identity $\text{Re}[\text{Tr}(\mathbf{A}\mathbf{A}^H)] = \text{Tr}(\mathbf{A}\mathbf{A}^H)$. Furthermore, we have

$$\max_{j \notin I} \text{Re} \left\{ \text{Tr} \left[\left(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \right) \mathbf{z}^H \mathbf{B}_j \right] \right\} \leq N_a \max_{i \neq j} \text{Re}[\text{Tr}(\mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \mathbf{z}^H \mathbf{B}_j)] \quad (43)$$

$$\leq N_a \max_{i \neq j} \|\mathbf{B}_j^H \mathbf{B}_i\| \|\mathbf{s}_i \mathbf{z}^H \mathbf{B}_j\| \quad (44)$$

$$= N_a d\mu_{\mathbf{B}} \tau s_u \quad (45)$$

By the definition of τ

$$\max_{j \notin I} \text{Tr}\{\mathbf{B}_j^H \mathbf{z} \mathbf{z}^H \mathbf{B}_j\} = \max_{j \notin I} \|\mathbf{B}_j^H \mathbf{z}\|_2^2 \leq \tau^2 \quad (46)$$

From derivations above, we obtain

$$\max_{j \notin I} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 \leq \rho_0 M N_a^2 (d\mu_{\mathbf{B}})^2 s_u^2 + 2\sqrt{\rho_0 M} N_a d\mu_{\mathbf{B}} \tau s_u + \tau^2 \quad (47)$$

On the other hand, we have

$$\max_{j \in I} \text{Tr}\left\{\left(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i\right)\left(\sum_{i \in I} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j\right)\right\} = \max_{j \in I} \text{Tr}\{\mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j \mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j\} \quad (48)$$

$$+ \max_{j \in I} 2\text{Re}\{\text{Tr}\{(\mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j) \left(\sum_{i \in I \setminus \{j\}} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j\right)\}\} \quad (49)$$

$$+ \max_{j \in I} \text{Tr}\left\{\left(\sum_{i \in I \setminus \{j\}} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i\right)\left(\sum_{i \in I \setminus \{j\}} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j\right)\right\} \quad (50)$$

For each summation term above, the first term

$$\max_{j \in I} \text{Tr}\{\mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j \mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j\} = \max_{j \in I} \{\mathbf{s}_j^H (\mathbf{B}_j^H \mathbf{B}_j)^2 \mathbf{s}_j\} \quad (51)$$

$$\geq \lambda_{\min}\{(\mathbf{B}_j^H \mathbf{B}_j)^2\} \max_{j \in I} \{\|\mathbf{s}_j \mathbf{s}_j^H\|\} \quad (52)$$

$$= [1 - (d-1)\nu]^2 s_u^2 \quad (53)$$

in which Gershgorin circle theorem and Rayleigh-Ritz theorem have been used. By Gershgorin circle theorem and the property of eigenvalue, all the eigenvalues of $(\mathbf{B}_j^H \mathbf{B}_j)^2$ are in the range $[(1 - (d-1)\nu)^2, (1 + (d-1)\nu)^2]$. At the same time, just like the derivation when $j \notin I$, the second summation term can be bounded as

$$\max_{j \in I} 2\text{Re}\{\text{Tr}\{(\mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j) \left(\sum_{i \in I \setminus \{j\}} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j\right)\}\} \geq -2(N_a - 1)[1 + (d-1)\nu] d\mu_{\mathbf{B}} s_l^2 \quad (54)$$

And for the third term

$$\max_{j \in I} \text{Tr}\{(\sum_{i \in I \setminus \{j\}} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i)(\sum_{i \in I \setminus \{j\}} \mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j)\} \geq 0 \quad (55)$$

Besides, since

$$\max_{j \in I} \text{Re}\{\text{Tr}\{(\sum_{i \in I} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i) \mathbf{z}^H \mathbf{B}_j\}\} \quad (56)$$

$$= \max_{j \in I} \text{Re}\{\text{Tr}\{\sum_{i \in I \setminus \{j\}} \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i \mathbf{z}^H \mathbf{B}_j + \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j \mathbf{z}^H \mathbf{B}_j\}\} \quad (57)$$

$$\geq -(N_a - 1)d\mu_{\mathbf{B}}\tau s_l - [1 + (d - 1)\nu]\tau s_l \quad (58)$$

and

$$\max_{j \in I} \text{Tr}\{\mathbf{B}_j^H \mathbf{z} \mathbf{z}^H \mathbf{B}_j\} \geq 0 \quad (59)$$

we have

$$\max_{j \in I} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 \geq \rho_0 M [1 - (d - 1)\nu]^2 s_u^2 - 2\rho_0 M (N_a - 1) [1 + (d - 1)\nu] d\mu_{\mathbf{B}} s_l^2 \quad (60)$$

$$- 2\sqrt{\rho_0 M} \tau [(N_a - 1)d\mu_{\mathbf{B}} + [1 + (d - 1)\nu]] s_l \quad (61)$$

Then we have

$$\max_{j \in I} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 - \max_{j \notin I} \|\mathbf{B}_j^H \mathbf{y}\|_2^2 \geq \rho_0 M [1 - (d - 1)\nu]^2 s_u^2 - \tau^2 - 2\sqrt{\rho_0 M} N_a d\mu_{\mathbf{B}} \tau s_u \quad (62)$$

$$- \rho_0 M d\mu_{\mathbf{B}} \{2(N_a - 1)[1 + (d - 1)\nu] s_l^2 + N_a^2 d\mu_{\mathbf{B}} s_u^2\} \quad (63)$$

$$- 2\sqrt{\rho_0 M} \tau \{(N_a - 1)d\mu_{\mathbf{B}} + [1 + (d - 1)\nu]\} s_l \quad (64)$$

By (62), Lemma 1 is proved. ■

Lemma 2. Suppose \mathbf{u} is a MT -dimensional circular symmetric Gaussian random vector of zero mean and \mathbf{I}_{MT} covariance matrix, then

$$P\{\tilde{\tau} \geq \|\mathbf{B}_j^H \mathbf{u}\|_2\} \geq 1 - e^{-\varsigma^2} \sum_{k=0}^{d-1} \frac{(\varsigma^2)^k}{k!} \quad (65)$$

where $\varsigma = \tilde{\tau} / \sqrt{1 + (d - 1)\nu}$.

Proof of Lemma 2: Suppose vector $\tilde{\mathbf{u}} = \sqrt{2}\mathbf{u}$, then $\tilde{\mathbf{u}}$ satisfies the distribution of $\mathcal{CN}(0, 2\mathbf{I})$. By [27], $\|\tilde{\mathbf{u}}\|_2^2$ is a chi-squared random variable with $2d$ degrees of freedom, then probability

$$\Pr\{\|\mathbf{u}\|_2^2 \geq t^2\} = \Pr\{\|\tilde{\mathbf{u}}\|_2^2 \geq 2t^2\} = \frac{\Gamma(d, t^2)}{\Gamma(d)} \quad (66)$$

where series expansion of $\Gamma(a, z)$ in [28] gives that

$$\Gamma(d, t^2) = \frac{e^{-t^2}}{2} t^2 [(\sqrt{2}t)^{2d} + (2d-2)(\sqrt{2}t)^{2d-2} + \dots + (2d-2)!!(\sqrt{2}t)^2] \quad (67)$$

By using the double factorial notation: $n!! = \prod_{0 \leq i \leq \lfloor n/2 \rfloor} (n-2i)$, and $\Gamma(d) = (d-1)!$, we will obtain

$$\frac{\Gamma(d, t^2)}{\Gamma(d)} = e^{-t^2} \left[\frac{(t^2)^{d-1}}{(d-1)!} + \frac{(t^2)^{d-2}}{(d-2)!} + \frac{(t^2)^{d-3}}{(d-3)!} + \dots + \frac{(t^2)^0}{0!} \right] \quad (68)$$

$$= e^{-t^2} \sum_{k=0}^{d-1} \frac{(t^2)^k}{k!} \quad (69)$$

which yields the lemma.

Consider the event $\tilde{\tau} \geq \|\mathbf{B}_j^H \mathbf{z}\|_2$, as in [20], $\mathbf{B}_j^H \mathbf{z}$ is a d -dimensional Gaussian random vector with zero mean and $\mathbf{B}_j^H \mathbf{B}_j$ covariance matrix. Therefore, vector $\mathbf{u} = (\mathbf{B}_j^H \mathbf{B}_j)^{-1/2} \mathbf{B}_j^H \mathbf{z}$ is also a d -dimensional circular symmetric Gaussian random vector of mean zero and covariance \mathbf{I}_d . Then by Lemma 2 and Result 1 it is easy to demonstrate that

$$\Pr\{\|\mathbf{B}_j^H \mathbf{z}\|_2^2 \leq \tilde{\tau}^2\} = \Pr\{\|(\mathbf{B}_j^H \mathbf{B}_j)^{1/2} \mathbf{u}\|_2^2 \leq \tilde{\tau}^2\} \quad (70)$$

$$\geq \Pr\{\|(\mathbf{B}_j^H \mathbf{B}_j)\| \cdot \|\mathbf{u}\|_2^2 \leq \tilde{\tau}^2\} \quad (71)$$

$$\geq \Pr\{\|\mathbf{u}\|_2^2 \leq \frac{\tilde{\tau}^2}{1 + (d-1)\nu}\} \quad (72)$$

$$= 1 - e^{-\varsigma^2} \sum_{k=0}^{d-1} \frac{(\varsigma^2)^k}{k!} \quad (73)$$

where $\varsigma = \tilde{\tau} / \sqrt{1 + (d-1)\nu}$. ■

With lemmas given above, we can prove Theorem 1 next.

Proof of Theorem 1: By the same induction in proof of [20, Theorem 5], when (13) is satisfied which verifies (35), for each iteration of BOMP, an most likely active user will be selected. Therefore, the actual support of the nonzero blocks can be correctly confirmed.

Gathering all these K selected blocks to form a set, say \hat{I} , we have $|\hat{I}| = K$, and $I \subseteq \hat{I}$, then

$$\|\hat{\mathbf{s}}_{BOMP} - \mathbf{s}\|_2^2 = \|(\sqrt{\rho_0 M} \mathbf{B}_{\hat{I}})^\dagger \mathbf{y} - \mathbf{s}\|_2^2 \quad (74)$$

$$= \|(\sqrt{\rho_0 M} \mathbf{B}_{\hat{I}})^\dagger \mathbf{z}\|_2^2 \quad (75)$$

$$\leq (\rho_0 M)^{-1} \|(\mathbf{B}_{\hat{I}}^H \mathbf{B}_{\hat{I}})^{-1}\|^2 \sum_{j \in \hat{I}} \|\mathbf{B}_j^H \mathbf{z}\|_2^2 \quad (76)$$

$$\leq \frac{K \tau^2}{[1 - (d-1)\nu - (K-1)d\mu_{\mathbf{B}}]^2 \rho_0 M} \quad (77)$$

where $\tau = \max_{1 \leq j \leq N} \|\mathbf{B}_j^H \mathbf{z}\|_2$ and (33) in Result 1 are used. The theorem is thus established. ■

REFERENCES

- [1] R. E. Grinter and L. Palen, "Instant messaging in teen life," in *Proceedings of the ACM conference on Computer supported cooperative work*, pp. 21-30, 2002.
- [2] Z. Xiao, L. Guo, and J. Tracey, "Understanding Instant Messaging Traffic Characteristics," in *Distributed Computing Systems, IEEE International Conference on*, pp. 51-51, 2007.
- [3] X. He, P.P.C. Lee, L. Pan, et. al., "A panoramic view of 3G data/control-plane traffic: mobile device perspective." *NETWORKING 2012, Springer Berlin Heidelberg*, pp. 318-330, 2012.
- [4] M. Park, R. Heath Jr., and S. M. Nettles, "Improving throughput and fairness for MIMO ad hoc networks using antenna selection diversity," *IEEE Global Telecommunications Conference*, vol. 5, pp. 3363-3367, 2004.
- [5] C.K. Pan, Y. M. Cai, and Y.Y. Xu, "Channel-aware multi-user uplink transmission scheme for SIMO-OFDM systems," *Science in China Series F: Information Sciences*, vol. 52, no. 9, pp. 1678-1687, 2009.
- [6] R. Xie, H. Yin, Z. Wang and X. Chen, "A Novel Uplink Data Transmission Scheme For Small Packets In Massive MIMO System," to appear at *IEEE/CIC 2014 Symposium on Signal Processing for Communications (ICCC)*.
- [7] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, Feb. 2014.
- [8] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40-60, 2013.
- [9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590-3600, 2010.
- [10] H. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436-1449, 2013.
- [11] D. Donoho, "Compressed Sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.
- [12] E. J. Candés, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489-509, 2006.

- [13] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118-120, 2007.
- [14] E. J. Candés and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21-30, 2008.
- [15] M. F. Duarte and Y. C. Eldar, "Structured Compressed sensing: From Theory to Applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053-4085, 2011.
- [16] Y. Li, J. H. Winters, and N. T. Sollenberger, "MIMO-OFDM for Wireless Communications: Signal Detection with Enhanced Channel Estimation," *IEEE Transactions on Communications*, vol. 50, no. 9, pp. 1471-1477, 2002.
- [17] P. Patel and J. Holtzman, "Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading," *IEEE International Conference on Communications*, pp. 510-514, 1994.
- [18] E. Biglieri, A. Nordin and G. Taricco, "MIMO Doubly-Iterative Receivers: Pre- vs. Post-Cancellation Filtering," *IEEE Commun. Lett.*, vol. 9, no. 2, pp. 106-108, 2005.
- [19] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042-3054, 2010.
- [20] Z. B. Haim and Y. C. Eldar, "Near-Oracle Performance of Greedy Block-Sparse Estimation Techniques From Noisy Measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1032-1047, 2011.
- [21] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982-2001, 2010.
- [22] L. Applebaum, W.U. Bajwa, M.F. Duarte, and R. Calderbank, "Asynchronous code-division random access using convex optimization," *Physical Communication*, vol. 5, no. 2, pp. 129-147, 2012.
- [23] H. Li, R. Mao, L. Lai, and R. C. Qiu, "Compressed meter reading for delay-sensitive and secure load report in smart grid," in *IEEE Int. Conf. on Smart Grid Commun.*, pp. 114-119, 2010.
- [24] R. H. Y. Louie, W. Hardjawana, Y. Li and B. Vucetic, "Distributed Multiple-Access for Wireless Communications: Compressed Sensing with Multiple Antennas," *IEEE Globecom*, pp. 3622-3627, 2012.
- [25] J. Luo and D. Guo, "Neighbor Discovery in Wireless Networks Using Compressed Sensing with Reed-Muller Codes," in *IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 154-160, 2011.
- [26] P. Viswanath and D. Tse, "Sum Capacity of the Vector Gaussian Broadcast Channel and Uplink-Downlink Duality," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1912-1921, 2003.
- [27] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6nd edition, London, U.K.: Edward Arnold, vol. 1, 1994.
- [28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, New York, Dover, 1964.
- [29] J. Kuang, *Applied Inequalities*, 4nd edition, Hunan Education Press, Changsha, China, 1993.
- [30] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *42nd Asilomar Conf. on Signals, Systems and Computers*, pp. 581-587, 2008.
- [31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.